

## Chapter 468

# Spectral Analysis

## Introduction

This program calculates and displays the periodogram and spectrum of a time series. This is sometimes known as harmonic analysis or the frequency approach to time series analysis.

Suppose we believe that a time series,  $X_t$ , contains a periodic (cyclic) component. A natural model of the periodic component would be

$$X_t = R \cos(ft + d) + e_t$$

where

- $R$  is the amplitude of variation. Normally, the cosine varies between -1 and 1. Hence, if  $R$  is 6, then the term would vary between -6 and 6. The impact of the amplitude is in the size (height or magnitude) of the wave. The length of the wave is not influenced by the amplitude.
- $f$  is the frequency of periodic variation, measured in number of radians per unit time. This is the 'frequency' scale of the plots. If we divide  $2\pi$  by  $f$ , we get the corresponding *wavelength*. This is the 'wavelength' scale of the plots. The impact of the frequency is to change the length of a cycle. As  $f$  increases, the length of the cycle decreases. A model with  $f=2$  would have a cycle length equal to one-half the cycle length of a model with  $f=1$ .
- $d$  is the phase. Changing the phase causes a shift in the beginning of the cycle.
- $e_t$  is the random error (noise) of the series about the period component.
- $t$  is the time period number. Usually,  $t=1, 2, 3, \dots, N$ .

Since  $\cos(ft+d) = \cos(ft) \cos(d) - \sin(ft) \sin(d)$ , this model may be written in the alternative form

$$X_t = a \cos(ft) + b \sin(ft) + e_t$$

where  $a = R \cos(d)$  and  $b = -R \sin(d)$ .

This model is a multiple regression model with two independent variables. In this case, the independent variables are  $X1 = \cos(ft)$  and  $X2 = \sin(ft)$ . The regression coefficients are  $B1 = a$  and  $B2 = b$ . In practice, the variation in a time series may be modeled as the sum of several different individual waves occurring at different frequencies.

The generalization of this model to the sum of  $k$  frequencies may be written symbolically as

$$X_t = \sum_{j=1}^k R_j \cos(f_j t + d_j) + e_t$$

or, using the alternative form, as

$$X_t = \sum_{j=1}^k a_j \cos(f_j t) + \sum_{j=1}^k b_j \sin(f_j t) + e_t$$

## Spectral Analysis

Note that if the  $f_j$  were known constants, and we let  $W_{jr} = \cos(f_r t)$  and  $Z_{js} = \sin(f_s t)$ , then this could be rewritten in the usual multiple regression form:

$$X_t = \sum_{j=1}^k a_j W_{ij} + \sum_{j=1}^k b_j Z_{ij} + e_t$$

where the  $a$ 's and the  $b$ 's are the regression coefficients to be estimated. This is an example of a harmonic regression.

Fourier analysis is the study of approximating functions using the sum of sine and cosine terms. This sum is called the Fourier series representation of the function. Spectral analysis is identical to Fourier analysis except that instead of approximating a function, the sum of sine and cosine terms approximates a time series that includes a random component. Note that the coefficients (the  $a$ 's and  $b$ 's) may be estimated using multiple regression.

One question that arises is how to select the frequencies. The highest frequency that can be fit to the data is  $\pi$ . The lowest is one cycle for the whole length of series, which amounts to a frequency of  $2\pi/N$  ( $N$  is the length of the series). Hence, one popular choice of frequencies is to select the  $N/2$  frequencies given by

$$f_k = 2\pi k / N, \quad (k = 1, 2, \dots, N/2)$$

The  $k^{\text{th}}$  frequency is often referred to as the  $k^{\text{th}}$  harmonic.

This set of frequencies is particularly popular when working by hand because it results in certain simplifications due to well-known trigonometric identities. However, there is nothing in nature that says that a series will follow these rather than some other set. That is why the program lets you specify a range of frequencies.

In the analysis of variance, we study the partitioning of the total variation (sum of squares) given by

$$SST = \sum_{t=1}^N (X_t - \bar{X})^2$$

into the sum of squares for factor A, factor B, etc. Similarly, in spectral analysis we are interested in partitioning the total sum of squares into amounts associated with each frequency. It turns out that the sum of squares for a particular frequency,  $SS_k$ , is given by

$$SS_k = \frac{N}{2} (a_k^2 + b_k^2)$$

If we regard  $SS_k$  as the portion of the total sum of squares accounted for by frequencies in the range

$$f_k \pm \frac{\pi}{N},$$

we can draw a histogram so that the area of each bar is proportional  $SS_k$ . The height of the histogram would be

$$I(f_k) = \frac{N}{4\pi} (a_k^2 + b_k^2)$$

The plot of  $I(f)$  versus  $f$  is called the *periodogram*.

This definition of the periodogram equates the total sum of squares to the area under the periodogram.  $I(f)$  may be calculated directly from the data as

$$I(f_k) = \frac{\left[ \sum X_t \cos(2\pi kt / N) \right]^2 + \left[ \sum X_t \sin(2\pi kt / N) \right]^2}{N\pi}$$

The periodogram is sometimes calculated using the fast Fourier transform (FFT). This method is not used in this program for three reasons. First, the increase in speed of the FFT is not significant until  $N$  is greater than one

## Spectral Analysis

thousand. For series of the length we normally anticipate for our users, the FFT would provide little speed improvement.

Second, when using the FFT, the length of the series ( $N$ ) must be a power of 2 (2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, etc.). If  $N$  is not a power of 2, then enough zeros must be added to bring the length of the series to the next power of 2. Suppose the length of a particular series was 260. You would need to add 252 zeros to bring the length to 512. This could dramatically distort your results. (FFT users use various “windows” or “filters” to remove the effect of these zeros. Since we do not pad with zeros, we do not need these filters.)

Third, we can calculate the periodogram for any set of frequencies, not just the set given above. This is very useful when you want to investigate a particular range of frequencies.

The sample periodogram has been shown to have some poor statistical properties. Recently, techniques for spectral analysis have improved on the periodogram by smoothing it. The smoothed periodogram is an estimate of the *power spectral density* or simply the *spectral density* of the series. The smoothing used in this program is simply an  $m$ -term moving average of the periodogram. The value of  $m$  is specified as the Smoothing Length option. Practitioners suggest that a value of  $m$  near  $N/40$  is reasonable. A large value of  $m$  may make the graph too smooth while a value too small may include spurious peaks.

Spectral analysis offers an interesting addition to other methods of time series analysis. For those who wish to find more out about it, we strongly recommend the book by C. Chatfield (1984). It offers a thorough, readable treatment of a difficult, but useful, subject.

## Data Structure

The data are entered in a single variable.

## Data Adjustment Options

### Remove Mean

Checking this option indicates that the series average should be subtracted from the data. This is almost always done.

### Remove Trend

Checking this option indicates that the least squares trend line should be subtracted from the data. This is sometimes done, although differencing is usually used to remove trends instead.

### Regular Differencing

This option lets you designate whether the original series, the first differences, or the second differences are analyzed. The first difference series,  $W$ , is calculated using the formula:

$$W_t = X_t - X_{t-1}$$

which may be written using the backshift operator,  $B$ , as:

$$W_t = (1 - B)X_t$$

The second difference series,  $Z$ , is the first difference of the  $W$  series. The formula is:

$$Z_t = W_t - W_{t-1}$$

which may be written using the backshift operator,  $B$ , as:

$$Z_t = (1 - B)^2 X_t$$

## Spectral Analysis

### Seasonal Differencing

This option lets you designate whether the original series, the first seasonal differences, or the second seasonal differences are analyzed. Assuming the number of seasons is  $s$ , the first seasonal difference series,  $W$ , is calculated using the formula:

$$W_t = X_t - X_{t-s}$$

which may be written using the backshift operator,  $B$ , as:

$$W_t = (I - B^s)X_t$$

The second seasonal difference series,  $Z$ , is the first seasonal difference of the  $W$  series. The formula is:

$$Z_t = W_t - W_{t-s}$$

which may be written using the backshift operator,  $B$ , as:

$$Z_t = (I - B^s)^2 X_t$$

---

## Missing Values

When missing values are found in the series, they are either replaced or omitted. The replacement value is the average of the nearest observation in the future and in the past or the nearest non-missing value in the past.

If you do not feel that this is a valid estimate of the missing value, you should manually enter a more reasonable estimate before using the algorithm. These missing value replacement methods are particularly poor for seasonal data. We recommend that you replace missing values manually before using the algorithm.

## Example 1 – Spectral Analysis

This section presents an example of how to do a spectral analysis of a time series. The Spots variable in the Sunspot dataset will be used.

### Setup

To run this example, complete the following steps:

**1 Open the Sunspot example dataset**

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Sunspot** and click **OK**.

**2 Specify the Spectral Analysis procedure options**

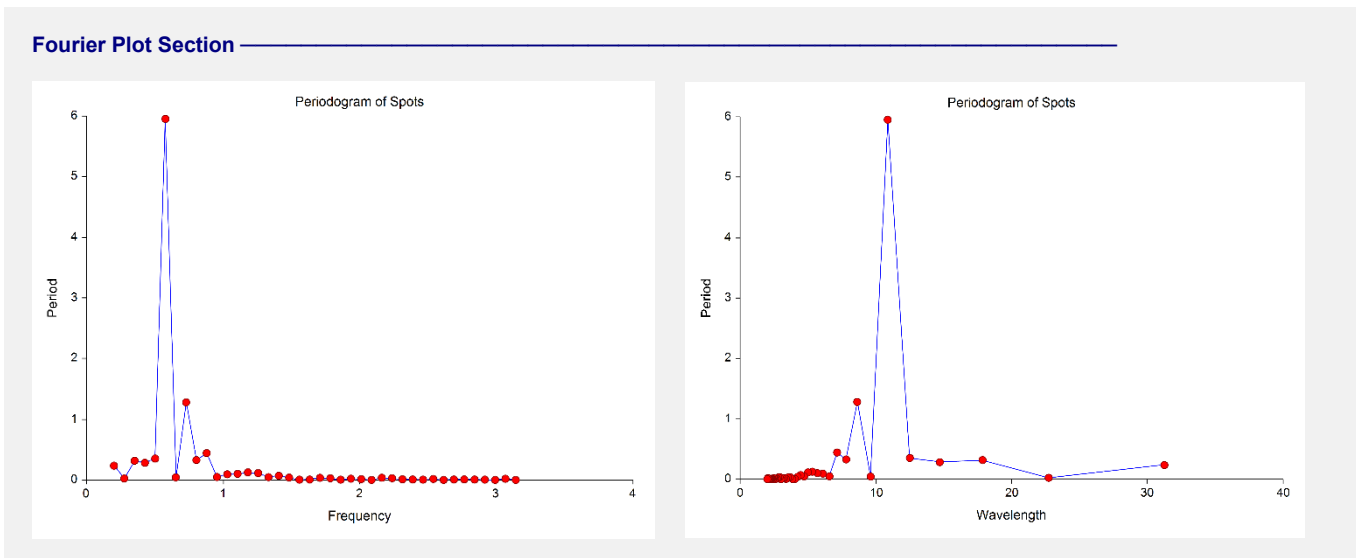
- Find and open the **Spectral Analysis** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1a** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

<u>Option</u>	<u>Value</u>
<b>Variables Tab</b>	
Time Series Variable .....	<b>Spots</b>

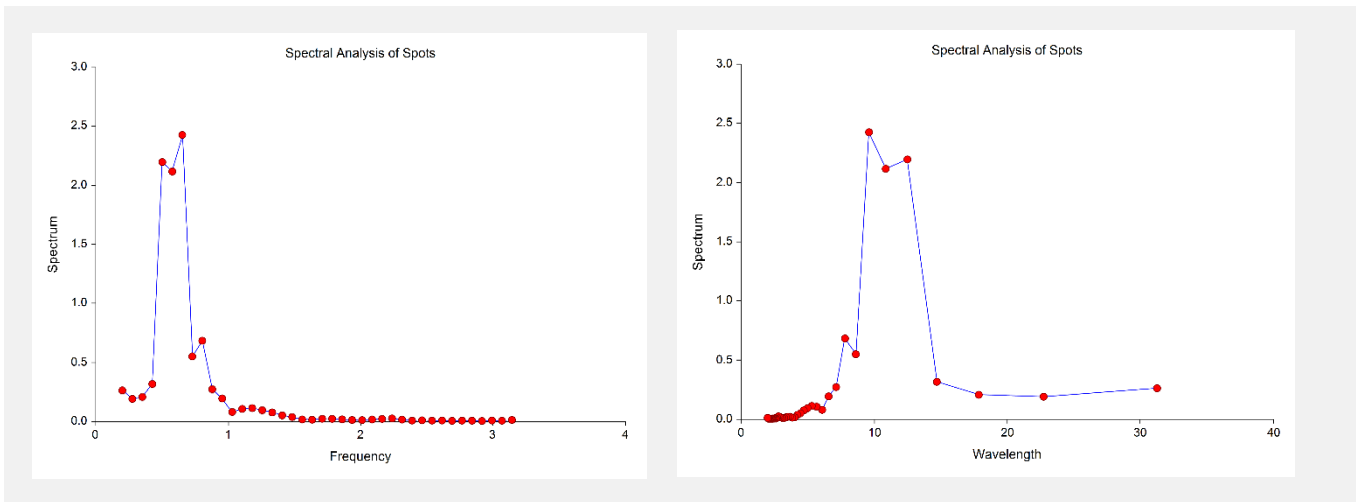
**3 Run the procedure**

- Click the **Run** button to perform the calculations and generate the output.

### Fourier Plot Section

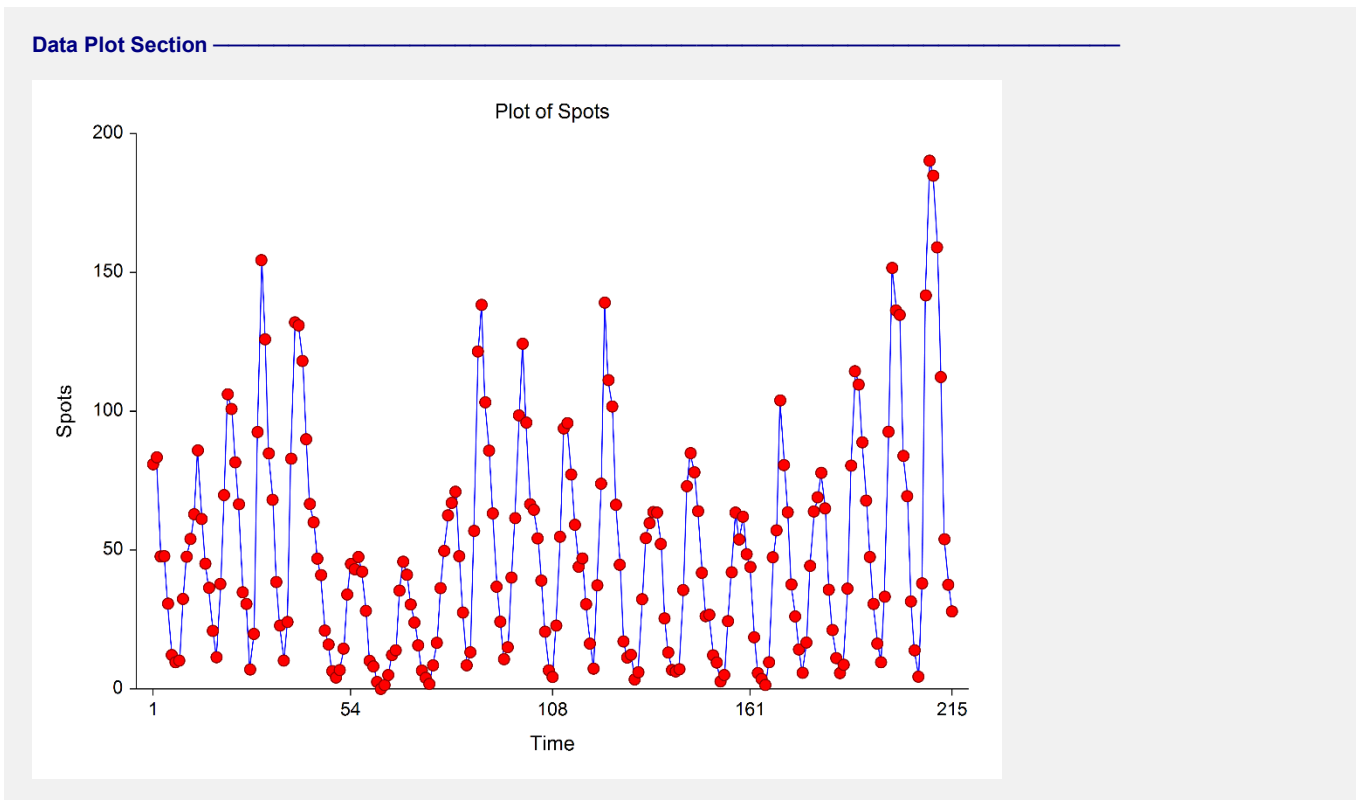


### Spectral Analysis



This section displays the periodogram and the spectrum plots with the frequency scale and the wavelength scale. Remember that the wavelength is in terms of the number of observations.

### Data Plot Section



This section displays a plot of the data values.

## Spectral Analysis

## Fourier Analysis Section

### Fourier Analysis of Spots (0,0,12,1,0)

Frequency	Wavelength	Period	Cosine(a's)	Sine(b's)	Spectrum
0.2010619	31.25	3200767	-75.89938	-425.8151	3590384
0.2764601	22.727427	324968.6	-13.48533	137.1568	2618775
0.3518584	17.85714	4330590	92.67065	-494.4972	2837492
0.4272566	14.70588	3856917	33.71187	-473.5963	4333876
0.5026549	12.5	4814120	298.3864	-438.5685	2.997943E+07
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

This section shows the values of the various components of the spectral analysis. The numbers in parentheses, (d,D,s,M,T), are defined as follows:

- d** is the regular differencing order.
- D** is the seasonal differencing order.
- s** is the number of seasons (ignored if D is 0).
- M** is 1 if the mean is subtracted, 0 otherwise.
- T** is 1 if the trend is subtracted, 0 otherwise.

## Fourier Plot Section

To complete this example, we rerun the analysis with the minimum wavelength set to 8 and the maximum wavelength set to 15. This appears to be the portion of the periodogram and spectrum that shows the most promise.

### 4 Specify additional Spectral Analysis procedure options

- The settings for this example are listed below and are stored in the **Example 1b** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

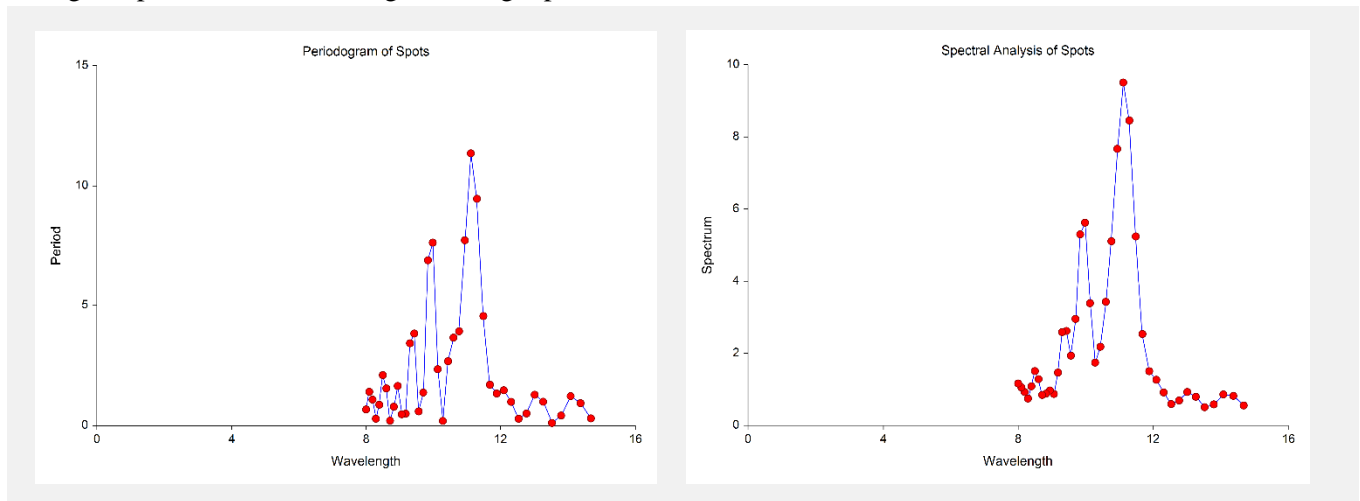
<u>Option</u>	<u>Value</u>
<b>Reports Tab</b>	
Minimum Wavelength .....	<b>8</b>
Maximum Wavelength .....	<b>15</b>

### 5 Re-Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

### Spectral Analysis

Doing this produces the following wavelength plots:



Now we can see the famous sunspot cycle of just over eleven years.